All 7 models - nb, C5.0, JRip, glm, kknn, nnet, svmlinear

jean wills

13/07/2020

THIS FILE IS TO SEE EFFECTS OF FEATURE SELECTION METHOD x5

## step 1a - need to get to “BM\_mini\_sc” (added 95% CI + numeric scaled)

library(plyr)  
library(dplyr)  
BM\_mini <- read.csv("/Users/jeanwills/Desktop/CKME136/1\_data/bank.csv", header=T, sep = ";", stringsAsFactors = T, na.strings = "NA")  
# Step 1 - NUMERIC DATA Cleaning - change numeric data outside the 2.5% and the 97.5% percentiles to this maximum/minimum value  
BM<-BM\_mini  
  
Lq\_bal<- quantile(BM$balance, probs=c(0.025))  
Hq\_bal<- quantile(BM$balance, probs=c(0.975))  
#Lq\_bal # -393  
#Hq\_bal # 8969  
Lq\_dur<- quantile(BM$duration, probs=c(0.025))  
Hq\_dur<- quantile(BM$duration, probs=c(0.975))  
#Lq\_dur # 19  
#Hq\_dur # 986  
Lq\_cam<- quantile(BM$campaign, probs=c(0.025))  
Hq\_cam<- quantile(BM$campaign, probs=c(0.975))  
#Lq\_cam # 1  
#Hq\_cam # 11  
Lq\_days<- quantile(BM$pdays, probs=c(0.025))  
Hq\_days<- quantile(BM$pdays, probs=c(0.975))  
#Lq\_days # -1  
#Hq\_days # 356  
Lq\_prv<- quantile(BM$previous, probs=c(0.025))  
Hq\_prv<- quantile(BM$previous, probs=c(0.975))  
#Lq\_prv # 0  
#Hq\_prv # 5  
  
BM$balance[BM$balance < Lq\_bal] <- Lq\_bal  
BM$balance[BM$balance > Hq\_bal] <- Hq\_bal  
  
BM$duration[BM$duration < Lq\_dur] <- Lq\_dur  
BM$duration[BM$duration > Hq\_dur] <- Hq\_dur  
  
BM$campaign[BM$campaign < Lq\_cam] <- Lq\_cam  
BM$campaign[BM$campaign > Hq\_cam] <- Hq\_cam  
  
BM$pdays[BM$pdays < Lq\_days] <- Lq\_days  
BM$pdays[BM$pdays > Hq\_days] <- Hq\_days  
  
BM$previous[BM$previous < Lq\_prv] <- Lq\_prv  
BM$previous[BM$previous > Hq\_prv] <- Hq\_prv  
  
# now make minor adjsutments  
# switch -1 -> 0 in 'pdays'  
BM$pdays<- ifelse(BM$pdays == -1, 0, BM$pdays)  
# switch duration in seconds to minutes for easier use  
BM$duration<- BM$duration/60  
  
# now have file BM ..

## step 1b - BM -> BM\_mini\_sc"

# KEEP: age-1, balance-6, day-10, duration-12, campaign-13, pdays-14, previous-15  
BMS<- BM  
# BMS<-BM\_mini  
normalize<- function(x) {  
 return ((x - min(x)) / (max(x) - min(x)))  
}  
BM\_s<- as.data.frame(lapply(BMS[,c(1,6,10,12:15)], normalize))  
# now recombine dataframes with the nominal components  
BM\_s$job<-BMS$job  
BM\_s$marital<-BMS$marital  
BM\_s$education<-BMS$education  
BM\_s$default<-BMS$default  
BM\_s$housing<-BMS$housing  
BM\_s$loan<-BMS$loan  
BM\_s$contact<-BMS$contact  
BM\_s$month<-BMS$month  
BM\_s$poutcome<-BMS$poutcome  
BM\_s$y<-BMS$y  
# convert  
BM\_mini\_sc<-BM\_s  
rm(BMS)  
rm(BM\_s)  
rm(BM)

## Step 1c - NOW DELETE THE ATTRIBUTES HERE AND CHANGE ANY data where there is [-17]

temp<-BM\_mini\_sc  
  
# FS1 Nnet LVQ - keep only 7: # 2, 4, 6, 7, 12, 14, 16 => delete 15, 13,11,10,9,8,5,3,1  
temp<- temp[-15]  
temp<- temp[-13]  
temp<- temp[-11]  
temp<- temp[-10]  
temp<- temp[-9]  
temp<- temp[-8]  
temp<- temp[-5]  
temp<- temp[-3]  
temp<- temp[-1]  
# then run SMotE models again  
BM<- temp  
summary(BM)

## balance duration pdays previous   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.04935 1st Qu.:0.0879 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.08940 Median :0.1717 Median :0.0000 Median :0.00000   
## Mean :0.17805 Mean :0.2450 Mean :0.1105 Mean :0.09082   
## 3rd Qu.:0.20006 3rd Qu.:0.3206 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## housing contact poutcome y   
## no :1962 cellular :2896 failure: 490 no :4000   
## yes:2559 telephone: 301 other : 197 yes: 521   
## unknown :1324 success: 129   
## unknown:3705   
##   
##

## step 2 - run training and test datasets FOR ALL CONFIGURATIONS

# str(BM\_mini\_sc)  
# rename dataset here:  
# BM<- BM\_mini\_sc  
set.seed(30)  
# get train and test datasets  
BM\_train\_index <- sample(nrow(BM), 0.7 \* nrow(BM))  
BM\_train<- BM[BM\_train\_index, ]  
BM\_test <- BM[-BM\_train\_index, ]  
# BM\_train\_labels <- BM[BM\_train\_index, 17]  
# BM\_test\_labels <- BM[-BM\_train\_index, 17]  
# note that we only balance the training sets  
# and leave test set as is.

## step 3 RUN SMOTE for training

# install.packages("DMwR")  
library(DMwR)  
library(grid)  
# this one is slow  
set.seed(50)  
smote\_train <- SMOTE(y ~ ., data = BM\_train)   
table(smote\_train$y)

##   
## no yes   
## 1440 1080

# now we name datasets that are used in models below   
# \*\*\*\*\*\*\*\*\*- change y from 17 to 8....  
x=smote\_train[,-8]  
trainsv=smote\_train  
train=smote\_train  
y=smote\_train$y  
test\_noy=BM\_test[,-8]  
test\_labels=BM\_test$y

# step 4 - now run models

## MODEL 1 NAIVE BAYES - uses caret

# FINAL PARAMETER VALUES USED WERE fL = 0, usekernel = TRUE and adjust = 1.  
library(klaR)  
library(caret)  
library(e1071)  
library(MASS)  
set.seed(520)  
nb\_grid <- expand.grid(fL= 0, usekernel= c("TRUE"), adjust=1)  
# nb\_grid <- expand.grid(fL= c(0,1), usekernel= c("TRUE", "FALSE"), adjust=c(0,1,2,3))  
nab\_mod<- train(x=x, y=y, method="nb", metric="ROC", tuneGrid=nb\_grid, trControl = trainControl(method="repeatedcv", number=10, repeats=10, classProbs=TRUE, summaryFunction = twoClassSummary, verboseIter=FALSE))  
# to see model results:  
# nab\_mod  
# predict output using test set - even though we did 10x10 cv above, we are also using 'validation' set of 30% of the data:  
nab\_pred<- predict(nab\_mod, test\_noy)  
s<-table(nab\_pred, test\_labels)  
# Confusion matrix  
print(confusionMatrix(s))

## Confusion Matrix and Statistics  
##   
## test\_labels  
## nab\_pred no yes  
## no 1084 88  
## yes 112 73  
##   
## Accuracy : 0.8526   
## 95% CI : (0.8326, 0.8711)  
## No Information Rate : 0.8814   
## P-Value [Acc > NIR] : 0.9994   
##   
## Kappa : 0.338   
##   
## Mcnemar's Test P-Value : 0.1039   
##   
## Sensitivity : 0.9064   
## Specificity : 0.4534   
## Pos Pred Value : 0.9249   
## Neg Pred Value : 0.3946   
## Prevalence : 0.8814   
## Detection Rate : 0.7988   
## Detection Prevalence : 0.8637   
## Balanced Accuracy : 0.6799   
##   
## 'Positive' Class : no   
##

### we can now run 10-fold on test dataset:

# copy in files you need and use test dataset only   
banking<-BM\_test  
  
# the other way is to run 10-fold on the test dataset and take the average of the (10 times) F1 measure   
folds<- createFolds(banking$y, k=10)  
 # create a function to do 10 folds of the data and run the statistics...  
 results <- lapply(folds, function(x) {  
 test<- banking[x,]  
 # \*\*\*\*\*\*\*\*\*- change y from 17 to ....  
 pred<- predict(nab\_mod, test[-8])  
 actual<- test$y  
 # PPV = TP/(TP+FP)  
 # pos<-posPredValue(table(pred, actual))  
 # I actually want: NPV= TN/(TN+FN) for precision of minority class  
 pr<-negPredValue(table(pred, actual))  
 # pr<-precision(table(pred, actual ))  
 # rec<- recall(table(pred, actual))  
 # i actually want specificity for recall of minority class  
 rec<- specificity(table(pred, actual))  
 F1<- 2 \* pr \* rec /(pr + rec)  
 return(F1)  
 })  
 #  
 # print(results)  
 value<-mean(unlist(results))  
 print(value)

## [1] 0.4147272

# print the average of the 10 F1 results for test set

## model 2 - C5.0 Decision Tree algorithm

strengths: numeric/nominal, easy interpretation weaknesses: biased splits, overfitting

# The final tuning parameters used for the original model were trials = 1, model = tree and winnow = FALSE.  
# Trials = an integer specifying the number of boosting iterations. A value of one indicates that a single model is used.  
# winnow: A logical: should predictor winnowing (i.e feature selection) be used.  
library(caret)  
set.seed(40)  
# c5\_grid <- expand.grid(trials=c(1,3,5), model = c("tree", "rules"), winnow = c(TRUE, FALSE))  
c5\_grid <- expand.grid(trials = 1, model = "tree", winnow = FALSE)  
c5\_ctrl<- trainControl(method="repeatedcv", number = 10, repeats=10, classProbs=TRUE, summaryFunction = twoClassSummary)  
c5\_mod<- train(x,y, method="C5.0", metric="ROC", tuneGrid=c5\_grid, trControl = c5\_ctrl, verbose=FALSE)  
c\_pred<- predict(c5\_mod, test\_noy)  
# Testing the result output  
s<-table(c\_pred, test\_labels)  
# Confusion matrix  
print(confusionMatrix(s))

## Confusion Matrix and Statistics  
##   
## test\_labels  
## c\_pred no yes  
## no 1044 61  
## yes 152 100  
##   
## Accuracy : 0.843   
## 95% CI : (0.8226, 0.862)  
## No Information Rate : 0.8814   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.3969   
##   
## Mcnemar's Test P-Value : 6.973e-10   
##   
## Sensitivity : 0.8729   
## Specificity : 0.6211   
## Pos Pred Value : 0.9448   
## Neg Pred Value : 0.3968   
## Prevalence : 0.8814   
## Detection Rate : 0.7693   
## Detection Prevalence : 0.8143   
## Balanced Accuracy : 0.7470   
##   
## 'Positive' Class : no   
##

### we can now run 10-fold on test dataset:

# copy in files you need and use test dataset only   
banking<-BM\_test  
  
# the other way is to run 10-fold on the test dataset and take the average of the (10 times) F1 measure   
folds<- createFolds(banking$y, k=10)  
 # create a function to do 10 folds of the data and run the statistics...  
 results <- lapply(folds, function(x) {  
 test<- banking[x,]  
 # \*\*\*\*\*\*\*\*\*- change y from 17 to ....  
 pred<- predict(c5\_mod, test[,-8])  
 actual<- test$y  
 # PPV = TP/(TP+FP)  
 # pos<-posPredValue(table(pred, actual))  
 # I actually want: NPV= TN/(TN+FN) for precision of minority class  
 pr<-negPredValue(table(pred, actual))  
 # pr<-precision(table(pred, actual ))  
 # rec<- recall(table(pred, actual))  
 # i actually want specificity for recall of minority class  
 rec<- specificity(table(pred, actual))  
 F1<- 2 \* pr \* rec /(pr + rec)  
 return(F1)  
 })  
 #  
 # print(results)  
 value<-mean(unlist(results))  
 print(value)

## [1] 0.4897212

# print the average of the 10 F1 results for test set

### to see model results:

#c5\_mod  
#c5\_mod$finalModel  
# another way to see the tree's decisions and best attributes:  
summary(c5\_mod)

##   
## Call:  
## (function (x, y, trials = 1, rules = FALSE, weights = NULL, control  
## = FALSE, sample = 0, earlyStopping = TRUE, label = "outcome", seed =  
## 2372L), verbose = FALSE)  
##   
##   
## C5.0 [Release 2.07 GPL Edition] Wed Jul 29 00:04:00 2020  
## -------------------------------  
##   
## Class specified by attribute `outcome'  
##   
## Read 2520 cases (8 attributes) from undefined.data  
##   
## Decision tree:  
##   
## poutcome = success: yes (247/15)  
## poutcome in {failure,other,unknown}:  
## :...duration <= 0.2088935:  
## :...previous <= 0.006551186: no (840/65)  
## : previous > 0.006551186:  
## : :...poutcome = unknown: yes (29)  
## : poutcome in {failure,other}:  
## : :...contact = unknown: yes (18/2)  
## : contact in {cellular,telephone}:  
## : :...balance <= 0.1405453: no (107/4)  
## : balance > 0.1405453:  
## : :...previous > 0.4984407: no (21/1)  
## : previous <= 0.4984407:  
## : :...pdays <= 0.4606742: yes (20/4)  
## : pdays > 0.4606742: no (29/7)  
## duration > 0.2088935:  
## :...previous <= 0:  
## :...duration > 0.762151: yes (156/19)  
## : duration <= 0.762151:  
## : :...contact = telephone:  
## : :...balance <= 0.3299725: yes (38/9)  
## : : balance > 0.3299725: no (7)  
## : contact = unknown:  
## : :...duration <= 0.4345503: no (121/7)  
## : : duration > 0.4345503:  
## : : :...housing = no: yes (15/5)  
## : : housing = yes:  
## : : :...duration <= 0.5049905: yes (9/2)  
## : : duration > 0.5049905: no (41/11)  
## : contact = cellular:  
## : :...duration <= 0.4146846:  
## : :...duration <= 0.2206128: yes (19/6)  
## : : duration > 0.2206128: no (215/54)  
## : duration > 0.4146846:  
## : :...housing = no: yes (68/28)  
## : housing = yes:  
## : :...duration <= 0.5542916: no (39/11)  
## : duration > 0.5542916:  
## : :...balance <= 0.04390088: no (7)  
## : balance > 0.04390088: yes (30/6)  
## previous > 0:  
## :...poutcome = unknown: yes (187)  
## poutcome in {failure,other}:  
## :...previous <= 0.1968537: yes (41)  
## previous > 0.1968537:  
## :...contact in {telephone,unknown}: yes (56/6)  
## contact = cellular:  
## :...housing = no: yes (60/20)  
## housing = yes:  
## :...duration > 0.6597725: yes (17/2)  
## duration <= 0.6597725:  
## :...previous <= 0.2076173: no (25/4)  
## previous > 0.2076173:  
## :...pdays <= 0.3477216: yes (8/1)  
## pdays > 0.3477216:  
## :...poutcome = failure:  
## :...previous <= 0.6506352: no (17/5)  
## : previous > 0.6506352: yes (14/4)  
## poutcome = other:  
## :...previous <= 0.8854208: yes (14/6)  
## previous > 0.8854208: no (5)  
##   
##   
## Evaluation on training data (2520 cases):  
##   
## Decision Tree   
## ----------------   
## Size Errors   
##   
## 32 304(12.1%) <<  
##   
##   
## (a) (b) <-classified as  
## ---- ----  
## 1305 135 (a): class no  
## 169 911 (b): class yes  
##   
##   
## Attribute usage:  
##   
## 100.00% poutcome  
## 90.20% duration  
## 90.20% previous  
## 40.48% contact  
## 14.64% housing  
## 10.28% balance  
## 4.25% pdays  
##   
##   
## Time: 0.0 secs

# we see the decision trees and....  
# ... and best attributes: see below at bottom....

## model 3 - JRip (rule learner)

# The final values used for the model were NumOpt = 10, NumFolds = 10 and MinWeights = 10.  
library(caret)  
library(RWeka)  
set.seed(50)  
# JR\_grid <- expand.grid(NumOpt=c(1,3,5,10), NumFolds=c(1,3,5,10),MinWeights=c(1,3,5,10))  
JR\_grid <- expand.grid(NumOpt=10, NumFolds=10, MinWeights=10)  
JR\_ctrl<- trainControl(method="repeatedcv", number = 10, repeats = 10, classProbs=TRUE, summaryFunction = twoClassSummary)  
JR\_mod<- train(x, y, method="JRip", metric="ROC", tuneGrid=JR\_grid, trControl = JR\_ctrl)  
JR\_pred<- predict(JR\_mod, test\_noy)  
s<-table(JR\_pred, test\_labels)  
# Confusion matrix  
print(confusionMatrix(s))

## Confusion Matrix and Statistics  
##   
## test\_labels  
## JR\_pred no yes  
## no 1040 63  
## yes 156 98  
##   
## Accuracy : 0.8386   
## 95% CI : (0.8179, 0.8578)  
## No Information Rate : 0.8814   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.3826   
##   
## Mcnemar's Test P-Value : 5.074e-10   
##   
## Sensitivity : 0.8696   
## Specificity : 0.6087   
## Pos Pred Value : 0.9429   
## Neg Pred Value : 0.3858   
## Prevalence : 0.8814   
## Detection Rate : 0.7664   
## Detection Prevalence : 0.8128   
## Balanced Accuracy : 0.7391   
##   
## 'Positive' Class : no   
##

### we can now run 10-fold on test dataset:

# copy in files you need and use test dataset only   
banking<-BM\_test  
  
# the other way is to run 10-fold on the test dataset and take the average of the (10 times) F1 measure   
folds<- createFolds(banking$y, k=10)  
 # create a function to do 10 folds of the data and run the statistics...  
 results <- lapply(folds, function(x) {  
 test<- banking[x,]  
 # \*\*\*\*\*\*\*\*\*- change y from 17 to ....  
 pred<- predict(JR\_mod, test[-8])  
 actual<- test$y  
 # PPV = TP/(TP+FP)  
 # pos<-posPredValue(table(pred, actual))  
 # I actually want: NPV= TN/(TN+FN) for precision of minority class  
 pr<-negPredValue(table(pred, actual))  
 # pr<-precision(table(pred, actual ))  
 # rec<- recall(table(pred, actual))  
 # i actually want specificity for recall of minority class  
 rec<- specificity(table(pred, actual))  
 F1<- 2 \* pr \* rec /(pr + rec)  
 return(F1)  
 })  
 #  
 # print(results)  
 value<-mean(unlist(results))  
 print(value)

## [1] 0.4671015

# print the average of the 10 F1 results for test set

### JRip rules compiled…

JR\_mod$finalModel

## JRIP rules:  
## ===========  
##   
## (pdays >= 0.000032) and (pdays <= 0.415474) => .outcome=yes (508.0/41.0)  
## (duration >= 0.767927) => .outcome=yes (177.0/23.0)  
## (duration >= 0.208652) and (pdays >= 0.421541) and (balance >= 0.097411) => .outcome=yes (136.0/21.0)  
## (duration >= 0.405067) and (housing = no) and (balance >= 0.065473) => .outcome=yes (70.0/17.0)  
## (duration >= 0.209335) and (contact = telephone) and (balance <= 0.305971) => .outcome=yes (44.0/10.0)  
## (contact = cellular) and (duration >= 0.549121) and (balance >= 0.046037) => .outcome=yes (40.0/9.0)  
## => .outcome=no (1545.0/226.0)  
##   
## Number of Rules : 7

# all for outcome yes

## model 4 - Logistic Regression

# no parameters  
library(caret)  
set.seed(520)  
log\_mod<- train(x, y, method="glm", metric="ROC", family=binomial(link="logit"), trControl = trainControl(method="repeatedcv", number=10, repeats=10, verboseIter=FALSE, classProbs = TRUE, summaryFunction = twoClassSummary))  
log\_pred<- predict(log\_mod, test\_noy)  
s<-table(log\_pred, test\_labels)  
# Confusion matrix  
print(confusionMatrix(s))

## Confusion Matrix and Statistics  
##   
## test\_labels  
## log\_pred no yes  
## no 1043 51  
## yes 153 110  
##   
## Accuracy : 0.8497   
## 95% CI : (0.8295, 0.8683)  
## No Information Rate : 0.8814   
## P-Value [Acc > NIR] : 0.9998   
##   
## Kappa : 0.4358   
##   
## Mcnemar's Test P-Value : 1.534e-12   
##   
## Sensitivity : 0.8721   
## Specificity : 0.6832   
## Pos Pred Value : 0.9534   
## Neg Pred Value : 0.4183   
## Prevalence : 0.8814   
## Detection Rate : 0.7686   
## Detection Prevalence : 0.8062   
## Balanced Accuracy : 0.7777   
##   
## 'Positive' Class : no   
##

### we can now run 10-fold on test dataset:

# copy in files you need and use test dataset only   
banking<-BM\_test  
  
# the other way is to run 10-fold on the test dataset and take the average of the (10 times) F1 measure   
folds<- createFolds(banking$y, k=10)  
 # create a function to do 10 folds of the data and run the statistics...  
 results <- lapply(folds, function(x) {  
 test<- banking[x,]  
 # \*\*\*\*\*\*\*\*\*- change y from 17 to ....  
 pred<- predict(log\_mod, test[-8])  
 actual<- test$y  
 # PPV = TP/(TP+FP)  
 # pos<-posPredValue(table(pred, actual))  
 # I actually want: NPV= TN/(TN+FN) for precision of minority class  
 pr<-negPredValue(table(pred, actual))  
 # pr<-precision(table(pred, actual ))  
 # rec<- recall(table(pred, actual))  
 # i actually want specificity for recall of minority class  
 rec<- specificity(table(pred, actual))  
 F1<- 2 \* pr \* rec /(pr + rec)  
 return(F1)  
 })  
 #  
 # print(results)  
 value<-mean(unlist(results))  
 print(value)

## [1] 0.5159

# print the average of the 10 F1 results for test set

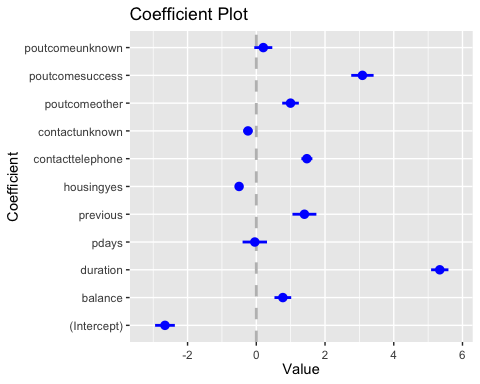
### model results

# to see model results:  
summary(log\_mod)

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9863 -0.6420 -0.4009 0.5786 2.4556   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.65736 0.28751 -9.243 < 2e-16 \*\*\*  
## balance 0.77248 0.24365 3.170 0.00152 \*\*   
## duration 5.33869 0.25226 21.163 < 2e-16 \*\*\*  
## pdays -0.04353 0.35546 -0.122 0.90253   
## previous 1.40018 0.34835 4.019 5.83e-05 \*\*\*  
## housingyes -0.50065 0.10870 -4.606 4.11e-06 \*\*\*  
## contacttelephone 1.47196 0.16015 9.191 < 2e-16 \*\*\*  
## contactunknown -0.24256 0.13326 -1.820 0.06874 .   
## poutcomeother 0.99614 0.24210 4.114 3.88e-05 \*\*\*  
## poutcomesuccess 3.08824 0.32563 9.484 < 2e-16 \*\*\*  
## poutcomeunknown 0.20375 0.26196 0.778 0.43668   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3441.9 on 2519 degrees of freedom  
## Residual deviance: 2225.8 on 2509 degrees of freedom  
## AIC: 2247.8  
##   
## Number of Fisher Scoring iterations: 5

### plot to see the most important attributes (those that “stand out” at far left or right)

require(coefplot)  
coefplot(log\_mod)



# to reinterpret data properly  
# invlogit<- function (x) {1/(1+exp(-x))}  
# invlogit(log\_mod$coefficients)  
# results of plot: duration, campaign, balance, poutcomesuccess, some months, , some jobs

## Model #5 - K-Nearest Neighbours

# The final values used for the model were kmax = 9, distance = 2 and kernel = optimal  
library(caret)  
library(lattice)  
library(ggplot2)  
knn\_ctrl<- trainControl(method="repeatedcv", number = 10, repeats = 10, classProbs=TRUE, summaryFunction = twoClassSummary)  
# knn\_grid<-expand.grid(kmax=c(5,7,9,13), distance=c(1,2,4,6), kernel=c("rectangular","rank","optimal"))  
knn\_grid<-expand.grid(kmax=9, distance=2, kernel="optimal")  
knn\_mod<- train(x=x,y=y, method="kknn", metric="ROC", tuneGrid=knn\_grid, trControl=knn\_ctrl, verbose=FALSE)  
knn\_pred<- predict(knn\_mod, test\_noy)  
# Testing the result output  
s<-table( knn\_pred, test\_labels)  
# Confusion matrix  
print(confusionMatrix(s))

## Confusion Matrix and Statistics  
##   
## test\_labels  
## knn\_pred no yes  
## no 1022 71  
## yes 174 90  
##   
## Accuracy : 0.8195   
## 95% CI : (0.7979, 0.8396)  
## No Information Rate : 0.8814   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.3239   
##   
## Mcnemar's Test P-Value : 7.195e-11   
##   
## Sensitivity : 0.8545   
## Specificity : 0.5590   
## Pos Pred Value : 0.9350   
## Neg Pred Value : 0.3409   
## Prevalence : 0.8814   
## Detection Rate : 0.7531   
## Detection Prevalence : 0.8055   
## Balanced Accuracy : 0.7068   
##   
## 'Positive' Class : no   
##

### we can now run 10-fold on test dataset:

# copy in files you need and use test dataset only   
banking<-BM\_test  
  
# the other way is to run 10-fold on the test dataset and take the average of the (10 times) F1 measure   
folds<- createFolds(banking$y, k=10)  
 # create a function to do 10 folds of the data and run the statistics...  
 results <- lapply(folds, function(x) {  
 test<- banking[x,]  
 # \*\*\*\*\*\*\*\*\*- change y from 17 to ....  
 pred<- predict(knn\_mod, test[-8])  
 actual<- test$y  
 # PPV = TP/(TP+FP)  
 # pos<-posPredValue(table(pred, actual))  
 # I actually want: NPV= TN/(TN+FN) for precision of minority class  
 pr<-negPredValue(table(pred, actual))  
 # pr<-precision(table(pred, actual ))  
 # rec<- recall(table(pred, actual))  
 # i actually want specificity for recall of minority class  
 rec<- specificity(table(pred, actual))  
 F1<- 2 \* pr \* rec /(pr + rec)  
 return(F1)  
 })  
 #  
 # print(results)  
 value<-mean(unlist(results))  
 print(value)

## [1] 0.4200348

# print the average of the 10 F1 results for test set

## NEURAL NET - model 6

Positive: can be used for classification or numeric prediction, makes few assumptions about the data (doesnt have to be normalized). Negative: SLOW..can overfit, ‘black box’.

# #1: The final parameter values used for the model were size = 16 and decay = 0.1.  
# Size is the number of units in hidden layer (nnet fit a single hidden layer neural network) and   
# decay is the regularization parameter to avoid over-fitting.  
require(mlbench)  
require(caret)  
require (nnet)  
nnctrl = trainControl(method="repeatedcv", number=10, repeats = 10, classProbs=TRUE, summaryFunction = twoClassSummary)  
# initially, create a grid list to find best parameters:  
# nn\_grid = expand.grid(size=c(1,4,8,16),decay=c(0,0.1,0.2,0.3,0.4))  
nn\_grid = expand.grid(size=16, decay=0.1)  
nn\_mod <- train(x=x, y=y, method="nnet", metric="ROC", trControl=nnctrl, tuneGrid=nn\_grid, trace=FALSE)  
nn\_pred<- predict(nn\_mod, test\_noy)  
# Testing the result output  
s<-table( nn\_pred, test\_labels)  
# Confusion matrix  
print(confusionMatrix(s))

## Confusion Matrix and Statistics  
##   
## test\_labels  
## nn\_pred no yes  
## no 1072 57  
## yes 124 104  
##   
## Accuracy : 0.8666   
## 95% CI : (0.8474, 0.8843)  
## No Information Rate : 0.8814   
## P-Value [Acc > NIR] : 0.9556   
##   
## Kappa : 0.4595   
##   
## Mcnemar's Test P-Value : 9.308e-07   
##   
## Sensitivity : 0.8963   
## Specificity : 0.6460   
## Pos Pred Value : 0.9495   
## Neg Pred Value : 0.4561   
## Prevalence : 0.8814   
## Detection Rate : 0.7900   
## Detection Prevalence : 0.8320   
## Balanced Accuracy : 0.7711   
##   
## 'Positive' Class : no   
##

### we can now run 10-fold on test dataset:

# copy in files you need and use test dataset only   
banking<-BM\_test  
  
# the other way is to run 10-fold on the test dataset and take the average of the (10 times) F1 measure   
folds<- createFolds(banking$y, k=10)  
 # create a function to do 10 folds of the data and run the statistics...  
 results <- lapply(folds, function(x) {  
 test<- banking[x,]  
 # \*\*\*\*\*\*\*\*\*- change y from 17 to ....  
 pred<- predict(nn\_mod, test[-8])  
 actual<- test$y  
 # PPV = TP/(TP+FP)  
 # pos<-posPredValue(table(pred, actual))  
 # I actually want: NPV= TN/(TN+FN) for precision of minority class  
 pr<-negPredValue(table(pred, actual))  
 # pr<-precision(table(pred, actual ))  
 # rec<- recall(table(pred, actual))  
 # i actually want specificity for recall of minority class  
 rec<- specificity(table(pred, actual))  
 F1<- 2 \* pr \* rec /(pr + rec)  
 return(F1)  
 })  
 #  
 # print(results)  
 value<-mean(unlist(results))  
 print(value)

## [1] 0.5333226

# print the average of the 10 F1 results for test set

## SVM model 7 - using caret package and linear kernel.

### Another supervised learning model. A SVM can be imagined as a surface thatc reates a boundary between points of data plotted in an n-space representing examples and their feature values. SVM creates a flat boundary called a hyperplane, which divides the space into homogeneous partitions on either side. This way, SVM combines nearest neighbour instance-based learning with linear regression modeling.

negatives: need to test various parameters and kernels to get best solution, can be slow, ’black box". positives: good for binary classification, classification/numeric prediction

# Linear (vanilla) kernel function.   
# The final values used for the model were Cost = 1 for classification.  
library(caret)  
library(kernlab)  
fitctrl<- trainControl(method="repeatedcv", number = 10, repeats = 10, classProbs=TRUE, summaryFunction = twoClassSummary)  
grid<-expand.grid(C=1)  
sv\_m<- train(y~., data=trainsv, method="svmLinear", metric="ROC", trControl=fitctrl, tunegrid=grid, trace=FALSE)  
sp<- predict(sv\_m, test\_noy)  
s<-table( sp, test\_labels)  
# Confusion matrix  
print(confusionMatrix(s))

## Confusion Matrix and Statistics  
##   
## test\_labels  
## sp no yes  
## no 1039 51  
## yes 157 110  
##   
## Accuracy : 0.8467   
## 95% CI : (0.8264, 0.8655)  
## No Information Rate : 0.8814   
## P-Value [Acc > NIR] : 0.9999   
##   
## Kappa : 0.4296   
##   
## Mcnemar's Test P-Value : 3.327e-13   
##   
## Sensitivity : 0.8687   
## Specificity : 0.6832   
## Pos Pred Value : 0.9532   
## Neg Pred Value : 0.4120   
## Prevalence : 0.8814   
## Detection Rate : 0.7657   
## Detection Prevalence : 0.8032   
## Balanced Accuracy : 0.7760   
##   
## 'Positive' Class : no   
##

### we can now run 10-fold on test dataset:

# copy in files you need and use test dataset only   
banking<-BM\_test  
  
# the other way is to run 10-fold on the test dataset and take the average of the (10 times) F1 measure   
folds<- createFolds(banking$y, k=10)  
 # create a function to do 10 folds of the data and run the statistics...  
 results <- lapply(folds, function(x) {  
 test<- banking[x,]  
 # \*\*\*\*\*\*\*\*\*- change y from 17 to ....  
 pred<- predict(sv\_m, test[-8])  
 actual<- test$y  
 # PPV = TP/(TP+FP)  
 # pos<-posPredValue(table(pred, actual))  
 # I actually want: NPV= TN/(TN+FN) for precision of minority class  
 pr<-negPredValue(table(pred, actual))  
 # pr<-precision(table(pred, actual ))  
 # rec<- recall(table(pred, actual))  
 # i actually want specificity for recall of minority class  
 rec<- specificity(table(pred, actual))  
 F1<- 2 \* pr \* rec /(pr + rec)  
 return(F1)  
 })  
 #  
 # print(results)  
 value<-mean(unlist(results))  
 print(value)

## [1] 0.5151625

# print the average of the 10 F1 results for test set